DOCUMENT RESUME

ED 432 604 TM 029 986

AUTHOR Buckendahl, Chad W.; Plake, Barbara S.; Impara, James C.
TITLE Setting Minimum Passing Scores on High-Stakes Assessments That Combine Selected and Constructed Response Formats.
PUB DATE 1999-04-00
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Standards; *Constructed Response; *Cutting Scores; Educational Assessment; Grade 3; Multiple Choice Tests; Primary Education; *Test Format; Test Items
IDENTIFIERS Angoff Methods; *High Stakes Tests; Paper Selection Method; *Standard Setting

ABSTRACT

Many school districts are developing assessments that incorporate both selected response and constructed response formats. Scores on these assessments can be used for a variety of purposes ranging from subject remediation to promotion decisions. These policy decisions are informed by recommendations for Minimum Passing Scores (MPSs) from standard setting studies. This paper presents a model for setting MPSs on mixed assessments using a combination of two standard setting methodologies. For multiple-choice items, an Angoff standard setting approach is used. For constructed response items, a paper selection strategy is used. The model is described in detail and an illustrative case is presented for a large metropolitan school district in the Midwest. (Contains 4 tables and 11 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 432 604

Setting Minimum Passing Scores on High-Stakes Assessments that Combine Selected and Constructed Response Formats

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Chad W. Buckendahl M.L.S.

Barbara S. Plake, Ph.D.

James C. Impara, Ph.D.

University of Nebraska-Lincoln

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Chad Buckendahl
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

A paper to be presented at the annual meeting
of the American Educational Research Association
Montreal, Quebec

April 25, 1999

TM029986

BEST COPY AVAILABLE

Abstract

Many school districts are developing assessments that incorporate both selected response and constructed response formats. Scores on these assessments can be used for a variety of purposes ranging from subject remediation to promotion decisions. These policy decisions are informed by recommendations for Minimum Passing Scores (MPSs) from standard setting studies. This paper presents a model for setting MPSs on mixed assessments using a combination of two standard setting methodologies. The model is described in detail and an illustrative case is presented for a large metropolitan school district in the Midwest.

Introduction

Test-centered methods for setting Minimum Passing Scores (MPSs), or cutscores, on selected response assessments have been well researched. The most prevalent method for setting MPSs on these types of assessments is the Angoff (1971) method (Sireci and Biskin, 1992). There is less congruence on methods for setting MPSs on tests using constructed response items. Some of the methods reported in the literature include the Judgmental Policy Capturing Method (Jaeger, 1994), the Dominant Profile Method (Plake, Hambleton, & Jaeger, 1997), an Integrated Judgment Method (Jaeger & Mills, 1998) and the Analytical Judgment Method (Plake & Hambleton, in press). Other reported methods include an extension of the Angoff method (Hambleton & Plake, 1995) and a paper selection approach (Hambleton, Jaeger, Mills, & Plake, in press), (Cross, Frary, Kelly, Small, & Impara, 1985).

Many school districts are developing assessments that blend selected response and constructed response formats. These mixed assessment approaches have been used to make high-stakes decisions about students (assignment to remedial or other "relooping-types" of educational programs or, in some cases, decisions about whether students will be permitted to graduate from high school).

These mixed assessments present challenges when making recommendations for the MPS as there is limited research on the effectiveness of standard setting methods with mixed assessments. The purpose of this paper is to present a model for setting MPSs on mixed assessments and to report on the results of one such application.

Standard Setting Model for Mixed Assessments

This model entails a mixture of standard setting approaches depending on the item type. For multiple-choice items, an Angoff standard setting approach is used. For constructed response items, a paper selection strategy is employed. Further, because these assessments are often long and complex, it

may not be feasible for a single panel to consider all the components of the test in a single day (an important consideration in a school district). Therefore, often it is necessary to hold the standard setting session over multiple days. This results in substantial cost for a school system.

The model proposed here entails subdividing both the panel and the assessment. Panels review a subpart of the assessment with overlapping assessment components being considered by each. An additional feature of this model is the provision of impact data on a periodic basis, showing the panelists the impact of the aggregate score as the panel members consider each of the assessment components (items or item sets) within their respective subparts. The purpose of this feature is to attempt to counteract the "Cascading Effect" (Plake, 1998) described by Linn and Shepard (1997) when aggregating MPS scores set on individual test components. Linn and Shepard, show that when test questions are not perfectly correlated, the result of aggregating MPS values on the individual test components is a final MPS that is more extreme than the panelists' likely anticipated from the impact data on the individual component parts.

In order to undertake this model for setting MPSs on mixed assessments (assessments combining selected and constructed response item types) the first step is to ascertain meaningful test component subparts to use as units for analysis (referred to as a "Rating Unit"). Often, especially for constructed response tasks, the student is asked to respond to several open-ended questions based on a common stimulus. For example, on a reading test, a student may be asked to a) summarize the main points of a story, b) answer questions about what happened in the story, c) make predictions about what might happen next in the story, and d) give a rationale for their predictions. Each of these components often has unique scoring guides or rubrics.

For the purposes of the standard setting activity, decisions need to be made about what constitutes a meaningful "rating unit". From the above

example, it is likely that parts a (summary) and b (answer questions) would be identified as unique rating units whereas parts c (prediction) and d (rationale) together would be deemed a single rating unit. This decision to combine parts c & d could be based in part (as shown in this example) on whether the parts are dependent or otherwise related in a logical way to each other. For this example, the subparts are dependent such that knowledge of the answer to part c (prediction) is needed to evaluate part d (rationale for prediction) even though the scoring may be distinct for these separate parts.

Once the rating units have been identified for the test, a decision is made about whether or not the test should be subdivided for standard setting. Our experience suggests that each rating unit, depending on the complexity of the task and the number of score points, will take 20 to 40 minutes for panelists to evaluate during standard setting. Therefore, in order to keep the standard setting activity confined to a 6 - 8 hour activity (including between 2 & 3 hours for orientation, training, and practice), it is unlikely that any one panel will be able to handle more than 8 - 10 rating units in a single day workshop.

If it is determined that a subdivision is necessary, the following guidelines are offered: a) the component parts should be balanced, as much as possible, in number of rating units and complexity of the tasks and b) at least one overlapping rating unit should be considered by both panels. We typically include one multiple choice and one constructed response rating unit as overlapping components. When the data are analyzed, the MPSs set on the overlapping components are compared to ascertain if the panels are providing estimates that are essentially on the same scale. If these values are sufficiently close (within a standard error), then we combine the results from the two panels directly to set the MPS on the full test. If these values are not close, some adjustment would be needed (this has never happened in our experience with this approach, however).

Methods and Procedure for an Illustrative Case

An illustration of this standard setting approach uses a Grade 3 Reading Assessment for a large metropolitan school district in the Midwest. The test is given to students at the end of the third grade year. The purpose of testing is to identify students who are reading below the Barely Proficient level (proficient students are those who can handle, with help reasonably provided by the classroom teacher, most third grade reading tasks).

Prior to convening the standard setting panel, anchor papers need to be selected by content experts and organized into sets, one set for each of the constructed response rating units. Anchor papers are student produced papers that "define" a given score point within a rating unit. These papers demonstrate typical responses and characteristics of each score point. Usually two anchor papers are chosen for each score point. Within sets, the anchor papers should be coded (so the panelists will not know the actual scored value of the paper) and organized in a random manner.

Panelists are initially convened in one large group for orientation and training that entails a discussion designed to elicit the knowledge, skills and abilities (KSAs) of the Minimally Competent Candidate (MCC). In many school settings other terms are used to identify the target student, such as the "Barely Proficient Reader" or the "Just Competent Student". Practice is undertaken for both selected response and constructed response type items. Typically, items used for practice are from a pilot or retired version of the test.

For practice in using selected response items, panelists make initial item performance estimates for the target student group and then engage in a guided discussion of their estimates. This activity is viewed as a continuation of the training where the KSAs articulated in the early discussion are applied to these particular practice test questions. Panelists are provided with item performance data and impact information (i.e., the percent

of students who would pass or fail using the cutscore based on the practice items). Following a discussion of the meaning of these data, panelists are given the opportunity to revise their initial (Round 1) estimates (i.e., make Round 2 item performance estimates).

For the constructed response practice, typically a simple rating unit is selected for training. Following presentation of the task and a description of the meaning of the score points, panelists are given a packet containing pre-coded, randomly ordered student work. The panelists' task is to select from this set of anchor papers the two that they feel are most indicative of the work of the target student. As was done when the multiple choice items were considered during training, panelists are asked to give their reasons for their selections in a group discussion prior to the presentation of actual student performance data. Following discussion of the meaning of this data, panelists have the opportunity to revise their paper selection decisions (Round 2 ratings).

When the panelists are convened in their small groups, they consider in sequential fashion, each of the rating units assigned to their group. It is desirable that the overlapping questions be considered in the same sequence in both groups to control for any order effects that might distort the comparability of the results across groups. Panelists make two rounds of student performance estimates for each rating unit. However, unlike what occurred during training, panelists do not engage in a discussion of their item performance estimates or paper selection decisions, between their Round 1 and Round 2 ratings. Actual student performance data and impact data are presented between Rounds 1 and 2 for each rating unit.

As panelists complete major components of the test (often a component of the test is broken into several rating units), the running total of MPSs derived for all the considered components is presented along with impact data on that component and all the other components collectively up to that point.

This information is shared with the panelists to help them understand the aggregate impact of the MPSs set up to this point in the standard setting process. This series of steps is repeated until the group has had an opportunity to consider all of the rating units assigned to that group. Because no group considers all the parts of the total assessment, the groups are not informed of the MPS derived for the total test.

Illustrative Case Description

As stated above, this case uses a standard setting approach for a Grade 3 Reading Assessment for a large metropolitan school district in the Midwest. The test seeks to identify students who are reading below the Barely Proficient level (proficient students are those who can handle, with help reasonably provided by the classroom teacher, most third grade reading tasks). The definition of Proficient provides a framework for developing an operational definition that can be applied to a measure of students' reading ability. In essence, the cutscore translates the verbal definition of proficiency to a definition on a test that differentiates between barely proficient readers and those who have not attained proficiency. In this illustrative case, special remedial programs will be made available to students who, based on their performance on this test and other relevant data, are deemed to be reading below the Barely Proficient level (i.e., students classified as Below Proficient).

The test includes both multiple choice items and performance tasks and is administered during the course of several days. Administration is in blocks lasting up to 45 minutes each for a total testing time of up to four hours. Slightly more than one-half the total-score points are from multiple choice items (77 out of 132). The performance tasks are scored on a variety of scales ranging from two to six point scales (consistent with the table of specifications). Rubrics, or scoring guides, are available for the performance tasks.

The test is composed of 7 parts. Part 1 consists of 19 multiple-choice questions and Part 7 assesses oral reading (accuracy, speed, and pronunciation that have an analytical scoring guide based on counts and time). The rest of the test is composed of a mixture of selected-response and constructed-response tasks based on short reading passages. An analysis of the components suggested subdividing the test into 16 rating units.

Because of the number of rating units, the test was divided into seven subparts for the Angoff standard setting study. The available teachers, then, were divided into two groups. Table 1 shows the breakdown of groups by the subparts of the test and the number of rating units within each subpart. Fifteen rating units were assigned to Group A (Subparts 1, 2, 4, 6 and 7) and thirteen were assigned to Group B (Subparts 1, 3, 5, and 7). A total of 29 panelists participated in the standard setting activity, 14 assigned to Group A and 15 assigned to Group B.

TABLE 1. Reading Assessment subparts and rating units by group.

Group	Subpart	Rating Units
A	1	1
	2	3
	4	3
	6	5
	7	3
B	1	1
	3	6
	5	3
	7	3

Both groups considered Parts 7 and 1 (in that order) first; then they proceeded in sequential order through the unique rating units assigned to their respective groups. After each subpart was completed, the panelists were shown

student data for that subpart and all of the subparts considered up to that point. After viewing the performance data, panelists were given an opportunity to revise their initial performance estimates.

Results

The first issue in examining the cutscore derived from this method is the extent that the ratings provided by the two groups of teachers who served as judges were equivalent. To test this equivalence, the cutscores for each group on Parts 1 and 7 (the overlapping Parts) of the test were examined. Results are shown below in Table 2. Group A had a final (round 2) cutscore for Part 1 of 10.29, and a final cutscore on Part 7 of 5.50. The comparable data from Group B were 9.73 for Part 1 and 4.47 for Part 7. These values are sufficiently close together (within one standard error) to justify treating the groups as equivalent. This judgment permits combining the data from the two groups without making any adjustment for one group being more lenient or severe than the other in terms of the overall group judgments.

TABLE 2. Round 2 data for overlapping subparts.

Group	Cutscore Subpart 1	S.D.	Cutscore Subpart 7	S.D.
A	10.29	1.52	5.50	0.54
B	9.73	1.68	4.47	0.89

The teachers provided performance estimates before and after being given actual performance data. The cutscores averaged across both groups from each round are shown in Table 3 below. Clearly the teachers were influenced by the data as the second round cutscore dropped from 66.12 to 63.91, slightly more than two points, between rounds one and two. The variation in cutscores also changed from round one to round two, increasing slightly from a standard

deviation of 6.73 in round one to 7.08 in round two. This change in variance is modest.

Table 3. Change in cutscore data between rounds 1 and 2.

Round	Cutscore	Standard Deviation	% below
1	66.12	6.73	18.20%
2	63.91	7.03	17.00%

The cutscores and associated ranges within which the final cutscore might be set as a result of using this standard setting method are shown in Table 4 below. If the cutscore was set at the average final value across the panels of teachers the cutscore would be 63.91 (SD = 7.08). The impact of this cutscore (rounded up to 64) would be that 17.00% of the third grade students in 1998 would be classified as being below proficient. If the cutscore were set at one standard deviation above the average cutscore (70.99), the impact would be that 21.60% of the third grade students would be classified as being Below Proficient.

Table 4. Round 2 cutscores and impact within 1 and 2 standard deviation range.

Range	Cutscore	Impact (% below)
2 SD Below	49.75	8.40%
1 SD Below	56.83	12.70%
Average	63.91	17.00%
1 SD Above	70.99	21.60%
2 SD Above	78.07	27.10%

In addition to the standard setting study, some validity data were collected to confirm procedural validity. These data were obtained from panelists' evaluations to ascertain levels of confidence in the MPS derived from this procedure. (See Giraud, Impara, & Buckendahl [in press] for a description of additional validation methods in standard setting.) These validity data are useful to establish quality and confidence in the results. Teachers' ratings suggested that panelists were comfortable with the process and confident that the MPS derived from the standard setting workshop was reasonable and appropriate.

Conclusion

Many school systems are adopting mixed assessment approaches for high stakes assessments. Most standard setting methods to date have focused exclusively on multiple-choice or constructed-response assessments. This paper presents an approach that was designed for use with mixed assessments and shows an illustration of an application with a Grade 3 Reading Assessment in a large metropolitan school system in the Midwest. The method has several strengths. It combines two approaches that have been shown to be effective with multiple-choice or constructed-response assessments; it can be applied to subparts of the test, allowing for a more efficient standard setting study; and the results were shown to be reasonable and appropriate when validity data were considered.

The application of a mixed assessment approach to standard setting warrants further study. As stated above, many schools are developing or using mixed assessment models for high stakes decisions (e.g. grade promotion, graduation). Because the illustration of the application in this paper was shown for a low stakes assessment, studies examining high stakes assessments would be useful. If this mixed assessment model of standard setting consistently suggests reasonable and defensible cutscores, school districts can add this method to the possible alternatives to use when setting cutscores on mixed assessment programs.

Acknowledgments

We would like to acknowledge Carla Noerrlinger, Virginia Brown, and Kathy Sullivan who were very helpful in designing the standard setting workshop and supplying feedback data to teachers. Also a special thanks to the teachers who participated in the study and the reading specialists who selected the anchor papers for the constructed response portions of the assessment. The success of the study was due, in large part, to their efforts.

References

Cross, L.H., Frary, R.B., Kelly, P.P., Small, R.C., & Impara, J.C. (1985). Establishing minimum standards for essays: Blind versus informed reviews. *Journal of Educational Measurement*, 22, 137-146.

Giraud, G., Impara, J.C., & Buckendahl, C. (in press). Alternative methods for standard setting in school districts. *Educational Assessment*.

Hambleton, R. K., & Plake, B.S. (1995) Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41-55.

Hambleton, R.K., Jaeger, R.M., Mills, C.N., & Plake, B.S. (in press). Handbook of methods for setting performance standards on complex performance assessments. Washington DC: Chief Council on State School Officers.

Jaeger, R.M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, 8, 15-40.

Jaeger, R.M., & Mills, C.N. (April, 1998). Integrated judgment method: A standard setting method designed for complex performance assessments with multiple performance categories, Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Linn, R.L., & Shepard, L. (July, 1997). Item-by-item standard setting: Misinterpretations of judge's intentions due to less than perfect item inter-correlations. Presentation at the Large Scale Assessment Conference, Colorado Springs, CO.

Plake, B.S. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education*, 11, 65-80.

Plake, B.S., & Hambleton, R.K. (in press). A standard setting method designed for complex performance assessments: Categorical assignments of student work. *Educational Assessment*.

Plake, B.S., Hambleton, R.K., & Jaeger, R.M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field test results. *Educational and Psychological Measurement*, 57, 400-412.

Sireci, S.A., & Biskin, B.H. (1992). Measurement practices in national licensing examination programs: A survey. *CLEAR Exam Review*, III(I), 21-25.